OcRFDet: Object-Centric Radiance Fields for Multi-View 3D Object Detection in Autonomous Driving

Mingqian Ji, Shanshan Zhang*, Jian Yang
PCA Lab, School of Computer Science and Engineering, Nanjing University of Science and Technology
{mingqianji,shanshan.zhang,csjyang}@njust.edu.cn

Abstract

Current multi-view 3D object detection methods typically transfer 2D features into 3D space using depth estimation or 3D position encoder, but in a fully data-driven and implicit manner, which limits the detection performance. Inspired by the success of radiance fields on 3D reconstruction, we assume they can be used to enhance the detector's ability of 3D geometry estimation. However, we observe a decline in detection performance, when we directly use them for 3D rendering as an auxiliary task. From our analysis, we find the performance drop is caused by the strong responses on the background when rendering the whole scene. To address this problem, we propose object-centric radiance fields, focusing on modeling foreground objects while discarding background noises. Specifically, we employ Object-centric Radiance Fields (OcRF) to enhance 3D voxel features via an auxiliary task of rendering foreground objects. We further use opacity - the side-product of rendering- to enhance the 2D foreground BEV features via Height-aware Opacity-based Attention (HOA), where attention maps at different height levels are generated separately via multiple networks in parallel. Extensive experiments on the nuScenes validation and test datasets demonstrate that our OcRFDet achieves superior performance, outperforming previous state-of-the-art methods with 57.2% mAP and 64.8% NDS on the nuScenes test benchmark. Code is available at https://github.com/Mingqj/OcRFDet.

1. Introduction

Multi-view 3D object detection plays a crucial role in autonomous driving. This task involves predicting object categories and their 3D coordinates based solely on multi-view RGB images. Currently, most detectors in this field project image features to the Bird's Eye View (BEV) space, and thus the main challenge lies in estimating 3D geometry within the BEV space using only RGB images. To address this problem, previous methods have made great efforts to learn effective geometric transformations from 2D to 3D space, but in a fully data-driven and implicit manner, which may limit performance improvement to some extent. Therefore, in this work, we aim to explore auxiliary tasks that are helpful to enhance the network's 3D geometry estimation capabilities, so as to facilitate multi-view 3D object detection.

*Corresponding author.

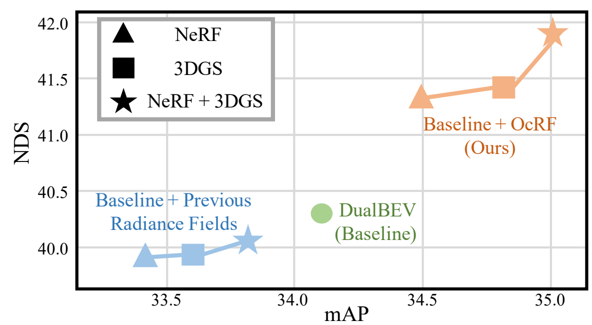

Figure 1. Comparison of previous radiance field methods directly applied to the baseline and our proposed OcRF applied to the baseline. Previous methods, whether using a single or joint field, consistently degrade performance, while our method enhances it, achieving greater improvements in the joint field setting.

Recently, radiance field methods, such as Neural Radiance Field (NeRF) [28] and 3D Gaussian Splatting (3DGS) [13], have shown strong geometry estimation capabilities in 3D reconstruction [27, 36, 41, 47] and simulation [45, 50, 52, 53] using knowledge-based geometric constraints and physical priors. Inspired by their success in occupancy prediction [11, 30, 31, 51] and BEV segmentation [2] tasks, we expect them to be also helpful for 3D object detection, which is yet under-explored in the community.

To explore the impact of radiance fields on the detection task, we directly apply the previous radiance fields and their rendering approach to an existing detector following the above methods, but we observe a decline in detection performance (Fig. 1) due to the strong responses on the background caused by rendering the whole scene (Fig. 2).

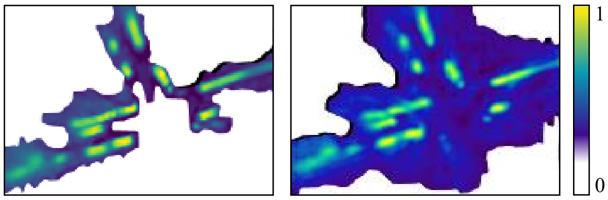

Figure 2. Comparison of BEV features before (left figure) and after (right figure) using previous radiance fields.

We argue that detection differs from the above perception tasks, in the way that it focuses on regions of interest instead of the entire scene, making the scene-centric rendering approach unsuitable for the detection task. Therefore, it is necessary to investigate how to make proper use of the radiance fields to enhance the detection performance.

In this paper, we propose a geometric feature enhancement approach based on the radiance fields to focus on modeling foreground objects of interest in both 3D voxel and 2D BEV space. For 3D voxel feature enhancement, we employ object-centric radiance fields (OcRF) to enhance foreground features via an auxiliary task of rendering 3D foreground objects. Specifically, we use the hybrid generalizable radiance fields combining NeRF and 3DGS to enhance 3D geometry estimation, and we develop an object-centric rendering approach tailored for the detection task by narrowing down the optimization goal from scene-level to object-level. This enables the radiance fields to more effectively and efficiently focus on modeling foreground objects of interest. It is worth noting that the rendering approach as an auxiliary head for training can be removed during model inference to reduce the network's computational burden. For 2D BEV feature enhancement, we propose height-aware opacity-based attention (HOA) to enhance foreground features based on geometric opacity, the side-product of rendering, which suggests the existence of foreground objects. Specifically, considering the variations in object distributions across different height levels (such as the presence of pedestrians and trees in lower-height levels and the absence of pedestrians in higher ones), we generate opacity-based attention maps at different height levels separately via multiple networks in parallel. These weights are applied to the BEV features, enabling the model to adaptively emphasize critical foreground features while suppressing invalid background noise.

Our contributions are summarized as follows:

- To our best knowledge, we are the first to apply radiance fields to multi-view 3D object detection in autonomous driving. We employ 3D foreground object rendering based on object-centric radiance fields as an auxiliary task to enhance geometry estimation within 3D voxel space.
- We further make use of opacity -the side-product of rendering- to enhance the foreground BEV features while suppressing irrelevant background noises. Considering the variation of object distributions at different height levels, we generate weights based on opacity at different height levels separately via multiple networks in parallel.
- Our method is evaluated on the nuScenes `validation` and `test` datasets [1], outperforming previous state-of-the-art methods.

2. Related Work

Since we focus on the task of multi-view 3D object detection and incorporate radiance fields to assist detection, we review recent works in multi-view 3D object detection, including depth-based and query-based detectors, and provide a brief overview of the radiance fields for perception.

2.1. Multi-View 3D Object Detection

Depth-based detectors. This type of detector estimates 3D geometry by learning depth distributions. BEVDet [10] follows LSS [34] to lift multi-view 2D camera features into 3D through implicit pixel-wise discrete depth prediction. BEVDepth [17] emphasizes the importance of depth in 3D geometry estimation and introduces explicit depth supervision. BEV-SAN [4] highlights the height information during view transformation. BEVNeXt [20] introduces conditional random fields to improve the accuracy of depth estimation. SA-BEV [49] introduces SA-BEVPool to enhance the geometry estimation of foreground objects by leveraging semantic segmentation as supervision. DualBEV [15] further extends this approach and proposes Prob-LSS to ignore invalid background features. Benefiting from the low computational cost of the depth-based methods, we choose this type of detector as our baseline, and introduce the radiance fields to enhance the geometry information of the foreground features, while emphasizing the height-aware representations in the radiance field.

Query-based detectors. This type of detector directly predicts 3D geometry in the BEV space by utilizing cross-attention to relate 2D image features with 3D queries. BEVFormer [18, 46] constructs a unified BEV space and designs spatial cross-attention to query the spatial 2D features. Based on BEVFormer, HeightFormer [43] introduces the height predictor to highlight the height information in BEV space. DETR3D [39] employs a sparse set of learnable reference points as 3D object queries to index 2D features. Based on this, Sparse4D [22] proposes a deformable 4D aggregation module to boost detection performance based on multi-dimensional features. SparseBEV [24] uses pillars instead of reference points as the 3D queries. QAF2D [12] introduces a pre-trained 2D detector to generate precise sparse 3D queries. RayDN [23] samples along camera rays to construct hard negative queries. OPEN [8] employs object-wise depth instead of typical pixel-wise depth for accurate detection. Our approach is also compatible with these methods,

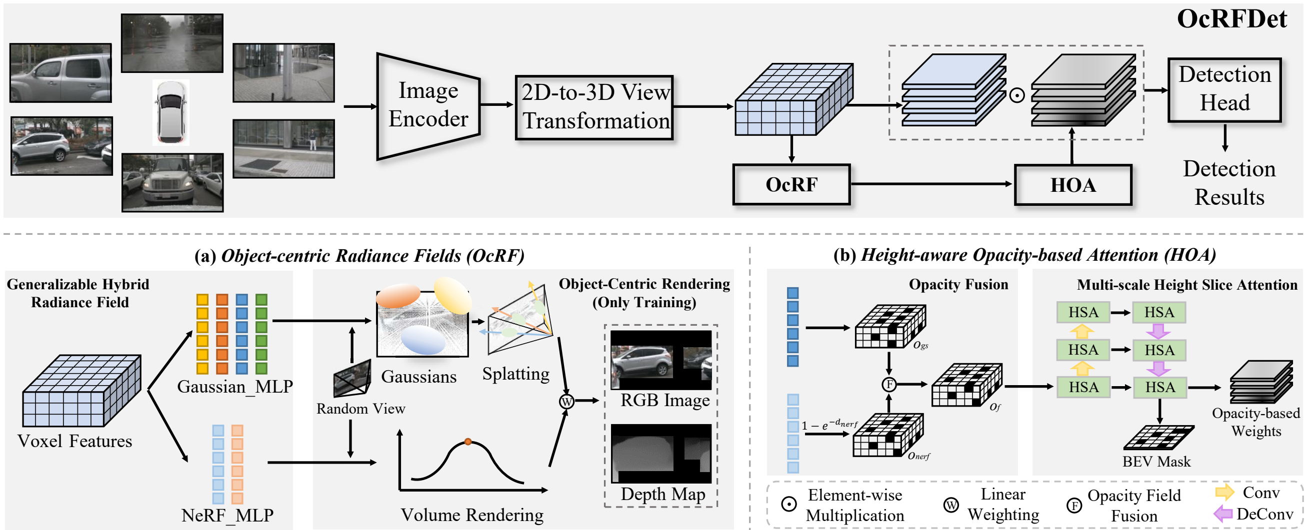

Figure 3. Overview of our method. It introduces the object-centric radiance fields (Sec. 3.2) and height-aware opacity-based attention (Sec. 3.3) within an existing detector to enhance 3D geometry estimation in both 3D voxel and 2D BEV space.

further enhancing the detection performance.

2.2. Radiance Field in Perception

Recently, radiance field methods have demonstrated strong geometry estimation capabilities in perception tasks such as occupancy prediction and BEV segmentation. In occupancy prediction, UniOcc [30] and RenderOcc [31] explicitly impose spatial geometric constraints through volumetric ray rendering. GaussianFormer [11] uses 3D Gaussian representations to model 3D scenes and performs occupancy prediction through Gaussian-to-voxel splatting. In BEV segmentation, GaussianBEV [2] uses 3D Gaussian representations for scene modeling and performs BEV segmentation via BEV splatting. The aforementioned scene-level radiance field methods are unsuitable for the detection task due to the introduction of irrelevant background information. To address this, we propose a novel approach tailored for the detection task.

3. Methodology

In this section, we first give an overview of our proposed multi-view 3D object detector, and then provide a detailed introduction to our proposed geometric enhancement method based on radiance fields.

3.1. Overview

We propose OcRFDet, a geometric feature enhancement approach based on the radiance fields, to enhance the detection network's 3D geometric modeling capabilities, building upon an existing detector. As illustrated in Fig. 3, our approach consists of two important modules: OcRF and HOA. In the following, we briefly describe the whole pipeline.

Image encoder. We use an image backbone with a feature pyramid network as the image encoder, which takes multi-view images as inputs, and encodes them into high-level image features.

2D-to-3D view transformation. This module transforms the image features from the multiple views to the bird's eye view. Following [17], we employ its Depth Net to predict depth probabilities, and a view transformer to lift the image features into pre-defined voxel space.

Object-centric radiance field. The OcRF module is designed to enhance the voxel features. We adopt hybrid generalizable radiance fields to improve generalization ability and 3D geometry estimation, and design an object-centric rendering approach tailored for detection task. Please refer to Sec. 3.2 for more details.

Height-aware opacity-based attention. The HOA module is designed to enhance the BEV feature. We introduce Height Slice Attention (HSA) into multi-scale opacity fields to obtain opacity-based attention maps at different height levels, and these weights are applied to the BEV features across different channel levels for enhancement. Please refer to Sec. 3.3 for more details.

Detection head. Based on the refined BEV features, we adopt a center-based 3D object detection head [48] to predict the object categories and 3D bounding boxes.

Optimization. We jointly optimize the detector and the radiance field branch end-to-end. The entire loss function is formulated as a summation of the detection loss, the rendering loss, and the BEV mask loss:

$$L = L_{det} + L_{render} + L_{mask}, \tag{1}$$

where the detection loss L_{det} is the same as the baseline detector. Please refer to Sec. 3.2.3 for rendering loss L_{render} and Sec. 3.3.2 for BEV mask loss L_{mask}.

3.2. Object-Centric Radiance Field

As shown in Fig. 3 (a), the object-centric radiance fields consist of hybrid generalizable radiance fields and an object-centric rendering approach. In the following, we provide a detailed explanation of each component.

3.2.1. Hybrid Generalizable Radiance Fields

Since traditional radiance fields rely on per-scene optimization, it is crucial to develop representations that can generalize across unseen scenes, particularly for the detection task. To this end, following [25] and [44], we employ multiple MLPs to decode 3D voxel features into randiance field representations, rather than typically encoding them from 2D RGB images. Moreover, we adopt a hybrid representation from 3DGS and NeRF to enhance generalization ability and improve geometric estimation.

Specifically, for the 3DGS representations, the attributes include position p_{gs}, scale s_{gs}, rotation r_{gs}, opacity o_{gs}, and color c_{gs}. For position, we directly scale the voxel coordinates to align with the actual perception range. Meanwhile, the remaining attributes are predicted through a Gaussian_MLP. The specific predictions are formulated as:

$$\begin{aligned} s_{gs} &= \text{Softplus}(S\_MLP(F_v)), \\ r_{gs} &= \text{Normalize}(R\_MLP(F_v)), \\ o_{gs} &= \text{Sigmoid}(O\_MLP(F_v)), \\ c_{gs} &= \text{Sigmoid}(C\_MLP(F_v)), \end{aligned} \tag{2}$$

where $S\_MLP(\cdot)$, $R\_MLP(\cdot)$, $O\_MLP(\cdot)$, and $C\_MLP(\cdot)$ each consists of a two-layer MLP. Besides, following [25], we apply specific activation functions to ensure the predicted values align with their physical ranges. For the NeRF representations, the attributes include density d_{nerf} and color weight w_{nerf}. Similarly, these attributes are predicted through a NeRF_MLP, and the network corresponding to each attribute consists of a two-layer MLP, and we apply specific activation functions to ensure the reasonability of predicted values.

As a result, we obtain the necessary attribute values, which are then used for rendering. These attribute values, predicted from high-level 3D voxel features, enable our model to generalize across diverse scenes.

3.2.2. Object-Centric Rendering

With the radiance field attributes available, we apply corresponding rendering techniques to generate RGB images and depth maps from specific viewpoints. Here, we randomly select one viewpoint from multiple views to render for a low computational cost. Specifically, for 3DGS attributes $\{p_{gs}, s_{gs}, r_{gs}, o_{gs}, c_{gs}\}$, we apply a degenerate depth-guided splatting regularization method [5], establishing a one-to-one correspondence between each 3DGS and a pixel to render the selected viewpoint. For the NeRF attributes $\{d_{nerf}, w_{nerf}\}$, we compute the opacity attribute o_{nerf}, as follows: $o_{nerf} = 1 - e^{-d_{nerf}}$, and the color attribute c_{nerf} is then obtained by applying the predicted color weights to the image plane features. After computing these NeRF attributes, we perform simple degenerate volume rendering, where only one point is sampled per ray to render the selected viewpoint, while producing the corresponding depth as the distance between the sampled point and the camera.

Unlike previous methods [2, 31] that optimize the entire scene, we only compute the loss values for the foreground regions based on the 2D labels, mapped from the 3D boxes to the image plane of the corresponding viewpoint. This object-centric rendering encourages both the Gaussian_MLP and NeRF_MLP to focus on modeling the foreground objects, thereby enhancing the quality of the voxel features. To maintain geometric consistency between the two radiance field networks, we further optimize the linear fusion of their rendered results:

$$\begin{aligned} I_f &= \alpha \cdot I_{gs} + \beta \cdot I_{nerf}, \\ D_f &= \alpha \cdot D_{gs} + \beta \cdot D_{nerf}, \end{aligned} \tag{3}$$

where $\{I_{gs}, I_{nerf}\}$ and $\{D_{gs}, D_{nerf}\}$ are the RGB images and depth maps rendered by 3DGS and NeRF, respectively; α and β are the learnable hyper-parameters, and $\alpha + \beta = 1$ $(\alpha, \beta > 0)$. Additionally, the scene-level optimization is employed as a warm-up phase to guide the initial training of the Gaussian_MLP and NeRF_MLP, and the whole rendering process is not involved during inference.

3.2.3. Rendering Loss

We optimize the radiance field branch with the mean squared error loss, structural similarity index measure loss [40], and $l1$ loss, as follows:

$$L_{render} = \lambda_{mse} \cdot L_{mse} + \lambda_{ssim} \cdot L_{ssim} + \lambda_{l1} \cdot L_{l1}, \tag{4}$$

where L_{mse} and L_{ssim} are typically used to render the RGB images, and L_{l1} is additionally introduced to enhance geometric modeling for depth map rendering; λ_{mse}, λ_{ssim} and λ_{l1} denote the weights for balancing loss items. Specifically, L_{mse} and L_{ssim} can be calculated as:

$$L_{mse} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\hat{I} \in I_p} [(I_i - \hat{I}_i) \odot M_i]^2, \tag{5}$$

$$L_{ssim} = \frac{1}{N} \sum_{i=1}^{N} \sum_{\hat{I} \in I_p} 1 - SSIM(I_i \odot M_i, \hat{I}_i \odot M_i), \tag{6}$$

where N is the number of samples; I represents the ground truth of RGB images; M is the 2D binary mask, where pixels inside the 2D ground truth boxes are set to 1 and those outside are set to 0; I_p contains the predicted results $\{I_{gs}, I_{nerf}, I_f\}$; $SSIM(\cdot)$ represents the SSIM function. Similarly, L_1 can be calculated as:

$$L_{l1} = \frac{1}{N}\sum_{i=1}^{N}\sum_{\hat{D}\in D_p} |(D_i - \hat{D}_i) \odot M_i|, \tag{7}$$

where D represents the ground truth of depth maps containing the cropped foreground objects; D_p contains the predicted results $\{D_{gs}, D_{nerf}, D_f\}$.

3.3. Height-Aware Opacity-based Attention

As shown in Fig. 3 (b), the HOA consists of an opacity fusion and multi-scale height slice attention. In the following, we provide a detailed explanation of each component.

3.3.1. Opacity Fusion

To enhance the geometric information of BEV features, we leverage the opacity field, a side-product of radiance field rendering, as it inherently encodes geometric knowledge within the radiance fields. We introduce an opacity fusion that fuses two geometric opacity fields $\{o_{gs}, o_{nerf}\}$ through a cross-attention mechanism, establishing fine-grained associations between the two fields and enhances the representation of geometric information:

$$o_f = CA(Q, K, V), \tag{8}$$

$$\begin{cases} Q = o_{nerf}, K = o_{gs}, V = o_{gs}, & \text{if } \alpha \leq \beta \\ Q = o_{gs}, K = o_{nerf}, V = o_{nerf}, & \text{if } \alpha > \beta \end{cases}, \tag{9}$$

where $CA(\cdot)$ denotes the cross-attention operation; α and β are the learnable hyper-parameters in Eq. 3. When the rendered images from some certain radiance field (3DGS or NeRF) have a higher weight (α or β), it reflects its superior rendering quality and high-quality geometry opacity. In such a case, the corresponding opacity field serves as the query Q in the cross-attention mechanism, while the other field acts as the key K and value V.

3.3.2. Multi-Scale Height Slice Attention

Based on the above fused opacity field, we further generate an attention map, which will be applied to the BEV features so as to highlight foreground objects.

Height slice attention. To refine the opacity field based on inner dependencies, we use self-attention at each layer. Inspired by previous works [4, 43] that highlight the importance of height information in BEV features, and consider the diverse distributions of objects at different heights, we employ multiple self-attention blocks in parallel, each for a certain height range. Specifically, the given opacity field is divided into k groups along height. Each of them is separately processed through MaxPooling, followed by a 1×1 convolution, converting the local geometric opacity information for each height interval into a corresponding geometric attention map. These attention maps are then concatenated in the order of height. Finally, a sigmoid activation function projects the attention map values to a range between 0 and 1, aligning them with their physical ranges.

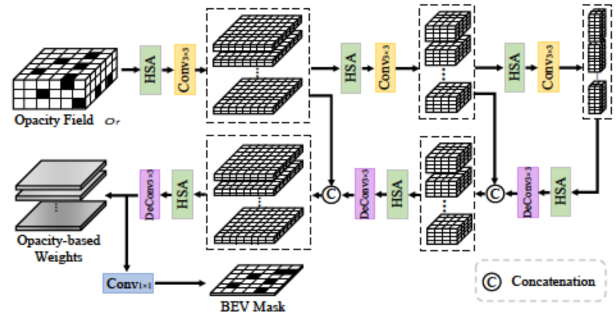

Figure 4. Illustration of multi-scale height slice attention.

Multi-scale integration. As shown in Fig. 4, a three-layer opacity pyramid is employed to integrate opacity information across multiple scales. Specifically, 3×3 convolution and deconvolution operations are applied to capture opacity at different scales. During scale transformation, HSA highlights the essential opacity information along the height dimension, while cascade connections retain original opacity details, enhancing the integration of multi-scale opacity.

BEV mask prediction. To suppress noises from the background regions, we add an auxiliary task of BEV mask prediction, following [14]. This mask is predicted by a 1×1 convolution and optimized by a combination of binary cross-entropy loss and dice loss [29], as follows:

$$L_{mask} = \lambda_{bce} \cdot L_{bce} + \lambda_{dice} \cdot L_{dice}, \tag{10}$$

where λ_{bce} and λ_{dice} denote the weights for balancing loss items.

4. Experiments

In this section, we will first introduce the dataset and evaluation metrics, followed by the implementation details. Then, we compare our method with the state-of-the-art methods. Finally, we will show the ablation studies and visualization.

4.1. Experimental Setup

Datasets and evaluation metrics. We evaluate our method on the nuScenes `validation` and `test` datasets [1]. We use the mean Average Precision (mAP) and nuScenes detection scores (NDS) to evaluate our detection results, where NDS is a comprehensive metric that combines the errors of object translation (mATE), scale (mASE), orientation (mAOE), velocity (mAVE), and attribute (mAAE).

Table 1. Comparisons with the state of the art on the nuScenes `validation` set. All methods use ResNet-50 [7] with FPN [21] as the image encoder, and the image resolution is set to 256 × 704.

Methods	Frames	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
BEVDet [10]	1	29.8	37.9	0.725	0.279	0.589	0.860	0.245
BEVFormer [18]	1	29.7	37.9	0.739	0.281	0.601	0.833	0.242
FB-BEV [19]	1	31.2	40.6	0.702	0.275	0.518	0.777	0.227
BEVDepth [17]	1	34.2	40.7	0.645	0.273	0.599	0.890	0.240
DualBEV [15]	1	35.2	42.5	0.640	0.271	0.542	0.838	0.216
OcRFDet (Ours)	1	**36.8**	**43.4**	0.620	0.269	0.519	0.848	0.242
BEVDet4D [9]	2	31.6	44.9	0.691	0.281	0.549	0.378	0.195
BEVFormer [18]	2	33.0	45.9	0.686	0.272	0.482	0.417	0.201
BEVDepth [17]	2	36.8	48.5	0.609	0.273	0.507	0.406	0.196
BEVStereo [16]	2	37.2	50.0	0.598	0.270	0.438	0.367	0.190
FB-BEV [19]	2	37.8	49.8	0.620	0.273	0.444	0.374	0.200
SA-BEV [49]	2	37.8	49.9	0.617	0.270	0.441	0.370	0.206
DualBEV [15]	2	38.0	50.4	0.612	0.259	0.403	0.370	0.207
OcRFDet (Ours)	2	**40.0**	**50.9**	0.582	0.277	0.518	0.375	0.216

Table 2. Comparisons with the state of the art on the nuScenes `test` set. By default, the image backbone is set as V2-99 [38], initialized from DD3D [32]; the image resolution is set to 640 × 1600. † indicate the use of ConvNeXt-B [26] as the image backbone. All methods consider the historical temporal information of 8 frames and employ 60-epoch training schemes for fair comparison.

Methods	mAP ↑	NDS ↑	mATE ↓	mASE ↓	mAOE ↓	mAVE ↓	mAAE ↓
SOLOFusion † [33]	54.0	61.9	0.453	0.257	0.376	0.276	0.148
SparseBEV [24]	54.3	62.7	0.502	0.244	0.324	0.251	0.126
StreamPETR [37]	55.0	63.6	0.479	0.239	0.317	0.241	0.119
DualBEV [15]	55.2	63.4	0.414	0.245	0.377	0.252	0.129
VideoBEV † [6]	55.4	62.9	0.457	0.249	0.381	0.266	0.132
Sparse4Dv2 [22]	55.6	63.8	0.462	0.238	0.328	0.264	0.115
BEVNeXt [20]	55.7	64.2	0.409	0.241	0.352	0.233	0.129
RecurrentBEV [3]	56.4	64.4	0.427	0.238	0.327	0.257	0.133
QAF2D [12]	56.6	64.2	0.461	0.240	0.326	0.261	0.121
RayDN [23]	56.5	64.5	0.461	0.241	0.322	0.239	0.114
OPEN [8]	56.7	64.4	0.456	0.244	0.325	0.240	0.129
OcRFDet (Ours)	**57.2**	**64.8**	0.421	0.253	0.331	0.248	0.130

Implementation details. We select DualBEV [15] as our baseline. The BEV grid size is set to 128 × 128 for the nuScenes `validation` dataset, and adjusted to 256 × 256 for the nuScenes `test` dataset. We integrate previous multiple frames following BEVDet4d [9]. Following 4DGS [42], the scene-level optimization is used as a warm-up during the first 2 epochs here. The hyper-parameter k in HSA is set to 4. We set the loss weights λ_{mse}, λ_{ssim}, λ_{l1}, λ_{bce}, λ_{dice} as 10, 1, 1, 10, and 10 to balance the loss values of auxiliary tasks and the detection task. Please refer to the supplementary material for more training details.

4.2. Main Results

We compare our OcRFDet with previous state-of-the-art methods on the nuScenes `validation` and `test` datasets. As illustrated in Tab. 1, for the nuScenes `validation` dataset, our OcRFDet achieves 36.8% mAP and 43.4% NDS performance with one single frame, which outperforms the state-of-the-art method (DualBEV) by 1.6 pp w.r.t. mAP and 0.9 pp w.r.t. NDS. When using 2 consecutive frames as inputs, OcRFDet achieves 40.0% mAP and 50.4% NDS performance, which outperforms the DualBEV by 2.0 pp w.r.t. mAP and 0.5 pp w.r.t. NDS. These results demonstrate the effectiveness of our method.

Furthermore, as illustrated in Tab. 2, for the nuScenes `test` dataset, our OcRFDet achieves 57.2% mAP and 64.8% NDS performance when using a longer 8-frame sequence as inputs, which outperforms DualBEV by 2.0 pp w.r.t. mAP and 1.1 pp w.r.t. NDS. Additionally, when compared to the state-of-the-art method (OPEN), which focuses on the temporal module, our OcRFDet maintains superiority and yields a new state-of-the-art result. These experiments further demonstrate the effectiveness of our method. More comparisons on the Waymo Open dataset [35] are provided in the supplemental material.

4.3. Ablation Studies

In this section, we conduct ablation studies to investigate the effectiveness of our OcRFDet. All method variants with one frame are trained with 20-epoch training strategy on

Figure 5. Visualization results of the 2D reference points.

Table 3. Comparisons of BEVFormer-tiny [18] as the baseline.

Methods	Backbone	BEV size	mAP ↑	NDS ↑
BEVFormer	ResNet-50	50×50	25.2	35.5
BEVFormer + Ours			**27.4**	**36.9**

Table 4. Ablations of each component.

Methods	mAP ↑	NDS ↑	Latency (ms) ↓
Baseline	34.1	40.3	**93.46**
+ OcRF	35.0	41.9	109.89
+ HOA	**35.5**	**42.2**	112.36

Table 5. Ablations of hybrid radiance fields.

3DGS	NeRF	mAP ↑	NDS ↑	SSIM ↑
✓		34.7	41.4	0.6803
	✓	34.5	41.3	0.6678
✓	✓	**35.0**	**41.9**	**0.7073**

Table 6. Ablations of depth rendering.

RGB Image	Depth Map	mAP ↑	NDS ↑
		34.1	40.3
✓		34.6	41.1
✓	✓	**35.0**	**41.9**

Table 7. Ablations of optimization goal.

Optimization Goal	mAP ↑	NDS ↑	SSIM ↑	
			Foreground	Entire scene
scene	33.8	40.1	0.6211	0.5731
object	34.7	41.5	0.6793	0.4825
scene + object	**35.0**	**41.9**	**0.7073**	**0.6134**

the nuScenes `validation` dataset. The latency tests are conducted on a single RTX 3090 GPU without TensorRT.

Generalization to query-based baseline. As shown in Tab. 3, when using BEVFormer-tiny [18] as our baseline on the nuScenes `validation` set, our method brings the improvements of 2.2 pp w.r.t. mAP and 1.4 pp w.r.t. NDS. Furthermore, as shown in Fig. 5, we visualize the 2D reference points with their weights in the BEVFormer encoder, and clearly find that our method makes the reference points of foreground have higher weights. These results demonstrate our method generalizes well to query-based detectors.

Effectiveness of each component. As shown in Tab. 4, compared with the baseline, our OcRF brings 0.9 pp w.r.t. mAP and 1.6 pp w.r.t. NDS improvements by enhancing 3D geometric features. On top of it, further adopting our HOA improves the performance by 0.5 pp w.r.t. mAP and 0.3 pp w.r.t. NDS by enhancing 2D BEV features. When combining both components, the results are improved significantly by 1.4 pp w.r.t. mAP and 1.9 pp w.r.t. NDS. These results demonstrate the effectiveness of each component. Additionally, we also test the latency of each component. The two components reveal an extra latency of 18.90 ms in total, with 16.43 ms from OcRF and 2.47 ms from HOA, which is minimal and thus friendly to applications.

Ablations of hybrid radiance fields. As shown in Tab. 5, using a single radiance field leads to a drop in detection performance and rendering quality. These suggest a strong correlation between geometric modeling capacity and detection accuracy. Hybrid radiance fields improve geometric consistency and structural completeness, enhancing feature quality and ultimately boosting detection performance.

Effectiveness of depth rendering. As shown in Tab. 6, compared with only rendering RGB images, the depth rendering brings 0.4 pp w.r.t. mAP and 0.8 w.r.t. NDS performance improvements by introducing explicit 3D geometric information. This result demonstrates the effectiveness of depth rendering in enhancing detection performance.

Effectiveness of rendering optimization goal. As shown in Tab. 7, optimizing the entire scene alone results in a 0.3% mAP and 0.2% NDS performance drop compared to the baseline. When optimizing the foreground objects, we observe significant performance improvements of 0.9 pp w.r.t. mAP, 1.4 pp w.r.t. NDS, and a rendering quality improvement of 5.82 pp w.r.t. SSIM. These results indicate our object-centric rendering is effective and more suitable for the detection task. Furthermore, using scene-level optimization as a warm-up phase further improves detection performance by 0.3 pp w.r.t. mAP and 0.4 pp w.r.t. NDS, indicating that this combination offers optimal results.

Ablations of opacity fusion. As shown in Tab. 8, when only using the opacity field from 3DGS to generate opacity-based attention, the mAP drops by 0.8 pp and the NDS by 0.8 pp; when only using it from NeRF, the mAP also drops by 0.9 pp and the NDS by 1.0 pp. We attribute this to the fact that the fused opacity incorporates richer geometric information compared to a single-source opacity, thereby providing greater benefits for detection. This experiment highlights the necessity of opacity fusion. Please refer to supplemental material for the ablations of fusion strategies.

Effectiveness of height slice attention. The results are shown in Tab. 9, when we remove our HSA, while retaining the multi-scale integration, the detection results drop w.r.t. both metrics. When applying a weight-sharing version of HSA by setting k to 1, the detection performance also drops, which shows the rationality of our HSA design. Moreover, while more height division (larger k) can capture more rich

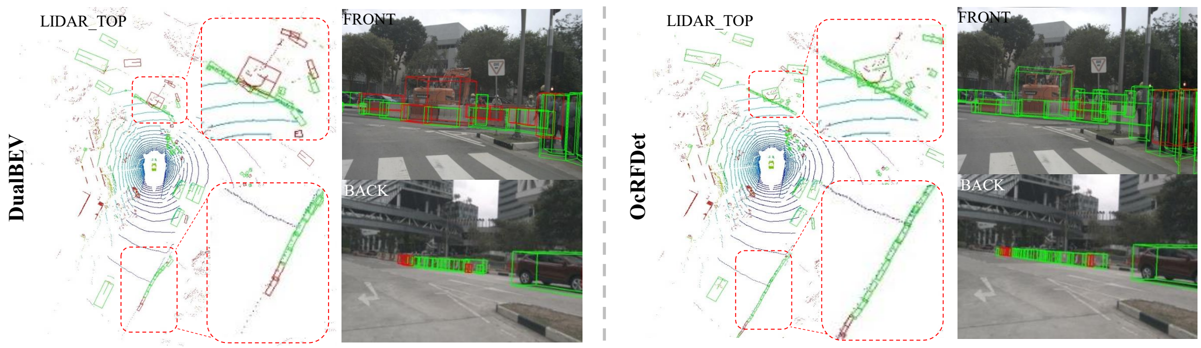

Figure 6. Qualitative detection results on images and the BEV space on the nuScenes `validation` set. We show the true positive boxes in green, and the incorrect prediction boxes in red. We use red rectangles to highlight the comparisons of ours with DualBEV.

Table 8. Ablations of opacity fusion.

Opacity Source	Fusion	mAP ↑	NDS ↑
3DGS	✗	34.7	41.4
NeRF	✗	34.6	41.2
3DGS + NeRF	✓	**35.5**	**42.2**

Table 9. Ablations of height slice attention.

Strategy	k	mAP ↑	NDS ↑	Latency (ms)
None	-	34.9	41.4	-
HSA	1	35.0	41.7	0.37
	2	35.3	41.9	0.71
	4	**35.5**	42.2	1.31
	6	**35.5**	**42.3**	2.04

height information, we find that setting $k = 4$ obtains the optimal balance between performance and latency. These results demonstrate the effectiveness and rationality of our HSA module. Please refer to supplemental material for the ablations of multi-scale height slice attention.

4.4. Visualization

Detection results. As shown in Fig. 6, we clearly find that our method shows fewer incorrect prediction boxes, including false and missed detections, especially for occluded and distant objects. These demonstrate that our approach is effective for detecting these challenging objects. More qualitative results are provided in the supplemental material.

Rendering results. As shown in Fig. 7, we present rendering results from three different views. We observe that our object-centric optimization produces clearer foreground objects compared to scene-level optimization, with high-quality rendered depth maps. These results demonstrate that our method's radiance field achieves a stronger geometric modeling capability for foreground objects.

BEV heatmap comparison. As shown in Fig. 8, our method using opacity-based attention maps exhibits higher-recall feature responses for distant objects compared to DualBEV, and the localization of feature responses is more aligned with the ground truth in crowded scenes. These indicate that our method effectively enhances features from a geometric perspective, further demonstrating the effectiveness of our approach.

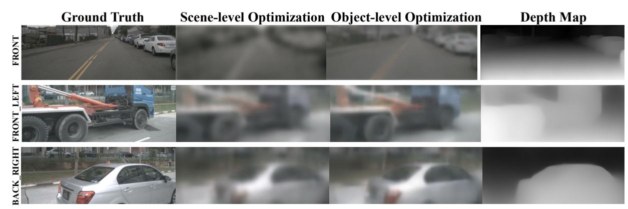

Figure 7. Visualization of rendering results.

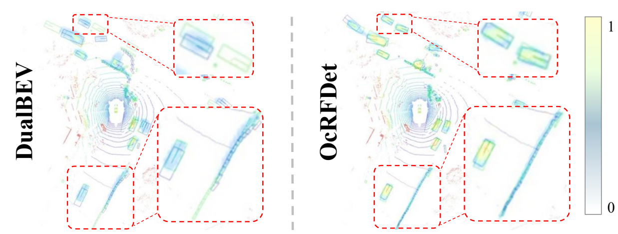

Figure 8. Comparison of BEV heatmaps. We show the ground truth boxes in **green**green, and the prediction boxes in blue.

5. Conclusion

In this paper, we propose OcRFDet, a geometric feature enhancement approach based on the radiance fields, designed to enhance the geometric information of foreground objects. Specifically, we employ object-centric radiance fields to enhance 3D voxel features via an auxiliary task of rendering foreground objects. Additionally, we propose a height-aware opacity-based attention to generate opacity-based attention maps to refine 2D BEV features. Extensive experiments demonstrate the effectiveness of our OcRFDet. Finally, we hope OcRFDet will further advance research on radiance fields for multi-view 3D object detection.

Acknowledgment

This research is supported by the National Natural Science Foundation of China (Grant No. 62322602), the Natural Science Foundation of Jiangsu Province, China (Grant No. BK20230033), and the National Natural Science Foundation of China (Grant No. 62172225).

References

[1] Holger Caesar, Varun Bankiti, Alex H Lang, Sourabh Vora, Venice Erin Liong, Qiang Xu, Anush Krishnan, Yu Pan, Giancarlo Baldan, and Oscar Beijbom. nuscenes: A multimodal dataset for autonomous driving. In *CVPR*, 2020. 2, 5

[2] Florian Chabot, Nicolas Granger, and Guillaume Lapouge. Gaussianbev: 3d gaussian representation meets perception models for bev segmentation. *arXiv preprint arXiv:2407.14108*, 2024. 1, 3, 4

[3] Ming Chang, Xishan Zhang, Rui Zhang, Zhipeng Zhao, Guanhua He, and Shaoli Liu. Recurrentbev: a long-term temporal fusion framework for multi-view 3d detection. In *ECCV*, 2024. 6

[4] Xiaowei Chi, Jiaming Liu, Ming Lu, Rongyu Zhang, Zhaoqing Wang, Yandong Guo, and Shanghang Zhang. Bev-san: Accurate bev 3d object detection via slice attention networks. In *CVPR*, 2023. 2, 5

[5] Jaeyoung Chung, Jeongtaek Oh, and Kyoung Mu Lee. Depth-regularized optimization for 3d gaussian splatting in few-shot images. In *CVPR*, 2024. 4

[6] Chunrui Han, Jinrong Yang, Jianjian Sun, Zheng Ge, Runpei Dong, Hongyu Zhou, Weixin Mao, Yuang Peng, and Xiangyu Zhang. Exploring recurrent long-term temporal fusion for multi-view 3d perception. *IEEE RA-L*, 2024. 6

[7] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *CVPR*, 2016. 6

[8] Jinghua Hou, Tong Wang, Xiaoqing Ye, Zhe Liu, Shi Gong, Xiao Tan, Errui Ding, Jingdong Wang, and Xiang Bai. Open: Object-wise position embedding for multi-view 3d object detection. In *ECCV*, 2024. 2, 6

[9] Junjie Huang and Guan Huang. Bevdet4d: Exploit temporal cues in multi-camera 3d object detection. *arXiv preprint arXiv:2203.17054*, 2022. 6

[10] Junjie Huang, Guan Huang, Zheng Zhu, Yun Ye, and Dalong Du. Bevdet: High-performance multi-camera 3d object detection in bird-eye-view. *arXiv preprint arXiv:2112.11790*, 2021. 2, 6

[11] Yuanhui Huang, Wenzhao Zheng, Yunpeng Zhang, Jie Zhou, and Jiwen Lu. Gaussianformer: Scene as gaussians for vision-based 3d semantic occupancy prediction. *arXiv preprint arXiv:2405.17429*, 2024. 1, 3

[12] Haoxuanye Ji, Pengpeng Liang, and Erkang Cheng. Enhancing 3d object detection with 2d detection-guided query anchors. In *CVPR*, 2024. 2, 6

[13] Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3d gaussian splatting for real-time radiance field rendering. *ACM Trans. Graph.*, 2023. 1

[14] Abhinav Kumar, Yuliang Guo, Xinyu Huang, Liu Ren, and Xiaoming Liu. Seabird: Segmentation in bird's view with dice loss improves monocular 3d detection of large objects. In *CVPR*, 2024. 5

[15] Peidong Li, Wancheng Shen, Qihao Huang, and Dixiao Cui. Dualbev: Cnn is all you need in view transformation. In *ECCV*, 2024. 2, 6

[16] Yinhao Li, Han Bao, Zheng Ge, Jinrong Yang, Jianjian Sun, and Zeming Li. Bevstereo: Enhancing depth estimation in multi-view 3d object detection with temporal stereo. In *AAAI*, 2023. 6

[17] Yinhao Li, Zheng Ge, Guanyi Yu, Jinrong Yang, Zengran Wang, Yukang Shi, Jianjian Sun, and Zeming Li. Bevdepth: Acquisition of reliable depth for multi-view 3d object detection. In *AAAI*, 2023. 2, 3, 6

[18] Zhiqi Li, Wenhai Wang, Hongyang Li, Enze Xie, Chonghao Sima, Tong Lu, Yu Qiao, and Jifeng Dai. Bevformer: Learning bird's-eye-view representation from multi-camera images via spatiotemporal transformers. In *ECCV*, 2022. 2, 6, 7

[19] Zhiqi Li, Zhiding Yu, Wenhai Wang, Anima Anandkumar, Tong Lu, and Jose M Alvarez. Fb-bev: Bev representation from forward-backward view transformations. In *ICCV*, 2023. 6

[20] Zhenxin Li, Shiyi Lan, Jose M Alvarez, and Zuxuan Wu. Bevnext: Reviving dense bev frameworks for 3d object detection. In *CVPR*, 2024. 2, 6

[21] Tsung-Yi Lin, Piotr Dollár, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *CVPR*, 2017. 6

[22] Xuewu Lin, Tianwei Lin, Zixiang Pei, Lichao Huang, and Zhizhong Su. Sparse4d: Multi-view 3d object detection with sparse spatial-temporal fusion. *arXiv preprint arXiv:2211.10581*, 2022. 2, 6

[23] Feng Liu, Tengteng Huang, Qianjing Zhang, Haotian Yao, Chi Zhang, Fang Wan, Qixiang Ye, and Yanzhao Zhou. Ray denoising: Depth-aware hard negative sampling for multi-view 3d object detection. In *ECCV*, 2024. 2, 6

[24] Haisong Liu, Yao Teng, Tao Lu, Haiguang Wang, and Limin Wang. Sparsebev: High-performance sparse 3d object detection from multi-camera videos. In *ICCV*, 2023. 2, 6

[25] Tianqi Liu, Guangcong Wang, Shoukang Hu, Liao Shen, Xinyi Ye, Yuhang Zang, Zhiguo Cao, Wei Li, and Ziwei Liu. Fast generalizable gaussian splatting reconstruction from multi-view stereo. In *ECCV*, 2024. 4

[26] Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *CVPR*, 2022. 6

[27] Ricardo Martin-Brualla, Noha Radwan, Mehdi SM Sajjadi, Jonathan T Barron, Alexey Dosovitskiy, and Daniel Duckworth. Nerf in the wild: Neural radiance fields for unconstrained photo collections. In *CVPR*, 2021. 1

[28] Ben Mildenhall, Pratul P Srinivasan, Matthew Tancik, Jonathan T Barron, Ravi Ramamoorthi, and Ren Ng. Nerf: Representing scenes as neural radiance fields for view synthesis. *Communications of the ACM*, 2021. 1

[29] Fausto Milletari, Nassir Navab, and Seyed-Ahmad Ahmadi. V-net: Fully convolutional neural networks for volumetric medical image segmentation. In *3DV*, 2016. 5

[30] Mingjie Pan, Li Liu, Jiaming Liu, Peixiang Huang, Longlong Wang, Shanghang Zhang, Shaoqing Xu, Zhiyi Lai, and Kuiyuan Yang. Uniocc: Unifying vision-centric 3d occupancy prediction with geometric and semantic rendering. *arXiv preprint arXiv:2306.09117*, 2023. 1, 3

[31] Mingjie Pan, Jiaming Liu, Renrui Zhang, Peixiang Huang, Xiaoqi Li, Hongwei Xie, Bing Wang, Li Liu, and Shanghang Zhang. Renderocc: Vision-centric 3d occupancy prediction with 2d rendering supervision. In *ICRA*, 2024. 1, 3, 4

[32] Dennis Park, Rares Ambrus, Vitor Guizilini, Jie Li, and Adrien Gaidon. Is pseudo-lidar needed for monocular 3d object detection? In *ICCV*, 2021. 6

[33] Jinhyung Park, Chenfeng Xu, Shijia Yang, Kurt Keutzer, Kris M Kitani, Masayoshi Tomizuka, and Wei Zhan. Time will tell: New outlooks and a baseline for temporal multi-view 3d object detection. In *ICLR*, 2022. 6

[34] Jonah Philion and Sanja Fidler. Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d. In *ECCV*, 2020. 2

[35] Pei Sun, Henrik Kretzschmar, Xerxes Dotiwalla, Aurelien Chouard, Vijaysai Patnaik, Paul Tsui, James Guo, Yin Zhou, Yuning Chai, Benjamin Caine, et al. Scalability in perception for autonomous driving: Waymo open dataset. In *CVPR*, 2020. 6

[36] Matthew Tancik, Vincent Casser, Xinchen Yan, Sabeek Pradhan, Ben Mildenhall, Pratul P Srinivasan, Jonathan T Barron, and Henrik Kretzschmar. Block-nerf: Scalable large scene neural view synthesis. In *CVPR*, 2022. 1

[37] Shihao Wang, Yingfei Liu, Tiancai Wang, Ying Li, and Xiangyu Zhang. Exploring object-centric temporal modeling for efficient multi-view 3d object detection. In *ICCV*, 2023. 6

[38] Tai Wang, Xinge Zhu, Jiangmiao Pang, and Dahua Lin. Fcos3d: Fully convolutional one-stage monocular 3d object detection. In *ICCV*, 2021. 6

[39] Yue Wang, Vitor Campagnolo Guizilini, Tianyuan Zhang, Yilun Wang, Hang Zhao, and Justin Solomon. Detr3d: 3d object detection from multi-view images via 3d-to-2d queries. In *RL*, 2022. 2

[40] Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE TIP*, 2004. 4

[41] Chenming Wu, Jiadai Sun, Zhelun Shen, and Liangjun Zhang. Mapnerf: Incorporating map priors into neural radiance fields for driving view simulation. In *IROS*, 2023. 1

[42] Guanjun Wu, Taoran Yi, Jiemin Fang, Lingxi Xie, Xiaopeng Zhang, Wei Wei, Wenyu Liu, Qi Tian, and Xinggang Wang. 4d gaussian splatting for real-time dynamic scene rendering. In *CVPR*, 2024. 6

[43] Yiming Wu, Ruixiang Li, Zequn Qin, Xinhai Zhao, and Xi Li. Heightformer: Explicit height modeling without extra data for camera-only 3d object detection in bird's eye view. *IEEE TIP*, 2024. 2, 5

[44] Chenfeng Xu, Bichen Wu, Ji Hou, Sam Tsai, Ruilong Li, Jialiang Wang, Wei Zhan, Zijian He, Peter Vajda, Kurt Keutzer, et al. Nerf-det: Learning geometry-aware volumetric representation for multi-view 3d object detection. In *ICCV*, 2023. 4

[45] Yunzhi Yan, Haotong Lin, Chenxu Zhou, Weijie Wang, Haiyang Sun, Kun Zhan, Xianpeng Lang, Xiaowei Zhou, and Sida Peng. Street gaussians for modeling dynamic urban scenes. *arXiv preprint arXiv:2401.01339*, 2024. 1

[46] Chenyu Yang, Yuntao Chen, Hao Tian, Chenxin Tao, Xizhou Zhu, Zhaoxiang Zhang, Gao Huang, Hongyang Li, Yu Qiao, Lewei Lu, et al. Bevformer v2: Adapting modern image backbones to bird's-eye-view recognition via perspective supervision. In *CVPR*, 2023. 2

[47] Jiawei Yang, Boris Ivanovic, Or Litany, Xinshuo Weng, Seung Wook Kim, Boyi Li, Tong Che, Danfei Xu, Sanja Fidler, Marco Pavone, et al. Emernerf: Emergent spatial-temporal scene decomposition via self-supervision. *arXiv preprint arXiv:2311.02077*, 2023. 1

[48] Tianwei Yin, Xingyi Zhou, and Philipp Krahenbuhl. Center-based 3d object detection and tracking. In *CVPR*, 2021. 3

[49] Jinqing Zhang, Yanan Zhang, Qingjie Liu, and Yunhong Wang. Sa-bev: Generating semantic-aware bird's-eye-view feature for multi-view 3d object detection. In *ICCV*, 2023. 2, 6

[50] Junge Zhang, Feihu Zhang, Shaochen Kuang, and Li Zhang. Nerf-lidar: Generating realistic lidar point clouds with neural radiance fields. In *AAAI*, 2024. 1

[51] Xiao Zhao, Bo Chen, Mingyang Sun, Dingkang Yang, Youxing Wang, Xukun Zhang, Mingcheng Li, Dongliang Kou, Xiaoyi Wei, and Lihua Zhang. Hybridocc: Nerf enhanced transformer-based multi-camera 3d occupancy prediction. *IEEE RA-L*, 2024. 1

[52] Hongyu Zhou, Jiahao Shao, Lu Xu, Dongfeng Bai, Weichao Qiu, Bingbing Liu, Yue Wang, Andreas Geiger, and Yiyi Liao. Hugs: Holistic urban 3d scene understanding via gaussian splatting. In *CVPR*, 2024. 1

[53] Xiaoyu Zhou, Zhiwei Lin, Xiaojun Shan, Yongtao Wang, Deqing Sun, and Ming-Hsuan Yang. Drivinggaussian: Composite gaussian splatting for surrounding dynamic autonomous driving scenes. In *CVPR*, 2024. 1